All SMILES Variational Autoencoder

Zaccary Alperstein
Variational.AI
zac@variational.ai

Artem Cherkasov
Vancouver Prostate Centre, UBC
acherkasov@prostatecentre.com

Jason Tyler Rolfe
D-Wave Systems
jrolfe@dwavesys.com

Abstract

Variational autoencoders (VAEs) defined over SMILES string (simplified molecular-input line-entry system) and graph-based representations of molecules promise to improve the optimization of molecular properties, thereby revolutionizing the pharmaceuticals and materials industries. However, these VAEs are hindered by the non-unique nature of SMILES strings. To efficiently pass messages along all paths through the molecular graph, we encode multiple SMILES strings of a single molecule using a set of stacked recurrent neural networks, pooling hidden representations of each atom between SMILES representations, and use attentional pooling to build a final fixed-length latent representation. By then decoding to a disjoint set of SMILES strings of the molecule, our All SMILES VAE learns an almost bijective mapping between molecules and latent representations near the high-probability-mass subspace of the prior. Our SMILES-derived but molecule-based latent representations significantly surpass the state-of-the-art in a variety of fully- and semi-supervised property regression and molecular property optimization tasks.

1 Introduction

The design of new pharmaceuticals, OLED materials, and photovoltaics all require optimization within the space of molecules [1]. While well-known algorithms such as gradient descent facilitate efficient optimization, they generally assume a continuous search space. In contrast, molecules correspond to graphs, with each node labeled by one of ninety-eight naturally occurring atoms, and each edge labeled as a single, double, or triple bond. Moreover, properties of interest are often sensitive to even small changes to the molecule [2], so their optimization is intrinsically difficult.

To solve this problem, previous works have trained a variational autoencoder (VAE) [3, 4] on SMILES string representations of molecules [5] to learn a decoder mapping from a Gaussian prior to the space of SMILES strings [6]. A sparse Gaussian process on molecular properties then facilitates Bayesian optimization of molecular properties within the latent space [6–9], or a neural network regressor from the latent space to molecular properties can be used to perform gradient descent on molecular properties with respect to the latent space [10–13].

SMILES, the simplified molecular-input line-entry system, defines a character string representation of a molecule by performing a depth-first pre-order traversal of a spanning tree of the molecular graph, emitting characters for each atom, bond, tree-traversal decision, and broken cycle [5]. The resulting character string corresponds to a flattening of a spanning tree of the molecular graph, as shown in Figure 1. The SMILES grammar is restrictive, and most strings over the appropriate character set do not correspond to well-defined molecules. Every molecule is represented by many well-formed SMILES strings, corresponding to all depth-first traversals of every spanning tree of the molecular graph. The distance between different SMILES strings of the same molecule can be much greater than that between SMILES strings from radically dissimilar molecules [11]. To address this difficulty, sequence-to-sequence transcoders [14] have been trained to map between different SMILES strings of a single molecule [15–17].

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

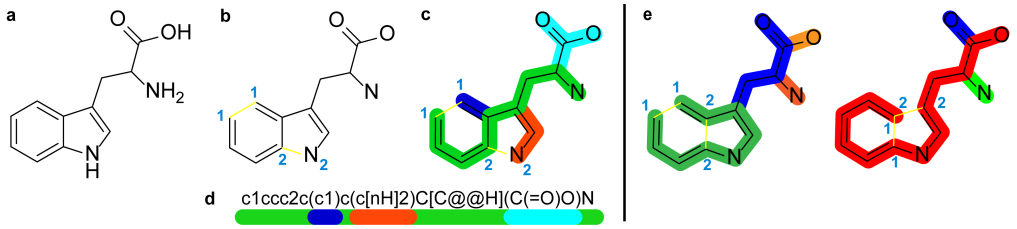


Figure 1: The molecular graph of the amino acid Tryptophan (a). To construct a SMILES string, all cycles are broken, forming a spanning tree (b); a depth-first traversal is selected (c); and this traversal is flattened (d). The beginning and end of intermediate branches in the traversal are denoted by (and) respective. The ends of broken cycles are indicated with matching digits. A small set of SMILES strings can cover all paths through a molecule (e).

We introduce the All SMILES VAE, which uses recurrent neural networks (RNNs) and atom-based pooling on multiple SMILES strings to implicitly perform efficient message passing along and amongst many flattened spanning trees of the molecular graph in parallel.

2 All SMILES variational autoencoder

A variational autoencoder (VAE) defines a generative model over an observed space x in terms of a prior distribution over a latent space $p(z)$ and a conditional likelihood of observed states given the latent configuration $p(x|z)$ [3, 4]. The true log-likelihood $\log p(x) = \log \int_z p(z)p(x|z)$, also known as the log-evidence in the Bayesian statistics literature, is intractable. The evidence lower bound (ELBO), based upon a variational approximation $q(z|x)$ to the posterior distribution, is maximized instead: $\mathcal{L} = \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] - \mathrm{KL}\left[q(z|x)||p(z)\right]$. The ELBO implicitly defines a stochastic autoencoder, where the approximating posterior $q(z|x)$ is the encoder and the conditional likelihood $p(x|z)$ is the decoder.

After the initial embedding, each layer of the All SMILES encoder takes representations of multiple distinct SMILES strings of the same molecule as input, and applies RNNs to them in parallel (specifically, gated recurrent units (GRUs) [18]), followed by atom-based harmonization and layer normalization, as shown in Figure 2. The RNNs implicitly realize a representative set of message passing pathways through the molecular graph, corresponding to the depth-first pre-order traversals of the spanning trees underlying the SMILES strings (Figure 1). To induce information flow amongst the union of the implicit SMILES pathways, for each atom, the atom harmonization step replaces the hidden representations from disparate SMILES strings with a pooled representation of that atom, as in Figure 6. The representations of the syntactical characters are unchanged by this harmonization.

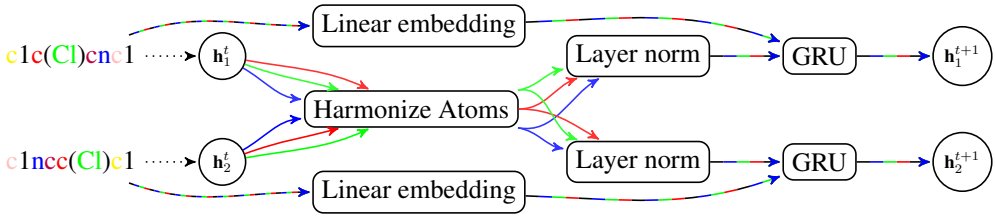


Figure 2: In each layer of the encoder after the initial BiGRU and linear transformation, hidden states corresponding to each atom are pooled across encodings of different SMILES strings for a common molecule, followed by layer norm and a GRU on each SMILES encoding independently.

The approximating posterior distills the resulting variable-length encodings into a fixed-length hierarchy of autoregressive Gaussian distributions [19]. The mean and log-variance of the first layer of the approximating posterior, $\mathbf{z}_1$, is parametrized by max-pooling the terminal hidden states of the final encoder GRUs, followed by batch renormalization [20] and a linear transformation, as shown in Figure 3. Succeeding hierarchical layers use Bahdanau-style attention [21]. The prior has a similar autoregressive structure, but uses fully connected layers of ReLUs in place of Bahdanau-style attention.

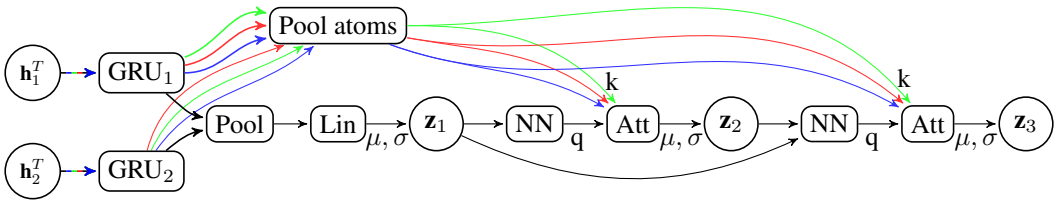


Figure 3: The approximating posterior is an autoregressive set of Gaussian distributions. The mean (μ) and log-variance ($\log \sigma^2$) of the first subset of latent variables $\mathbf{z}_1$ is a linear transformation of the max-pooled final hidden state of GRUs fed the encoder outputs. Succeeding subsets $\mathbf{z}_i$ are produced via Bahdanau-style attention with the pooled atom outputs of the GRUs as keys (k), and the query (q) computed by a neural network on $\mathbf{z}_{<i}$.

The decoder is a single-layer LSTM, for which the initial cell state is computed from the latent representation by a neural network, and a linear transformation of the latent representation is concatenated onto each input. It is trained with teacher forcing to reconstruct a set of SMILES strings disjoint from those provided to the encoder, but representing the same molecule.

Rather than apply a sparse Gaussian process to fixed latent representations to predict molecular properties [7–9, 11], the All SMILES VAE jointly trains property regressors with the generative model [6, 12]. We use linear regressors for the log octanol-water partition coefficient (logP) and molecular weight (MW), which have unbounded values; and logistic regressors for the quantitative estimate of drug-likeness (QED) [22] and twelve binary measures of toxicity [23, 24], which take values in $[0, 1]$. We then perform gradient-based optimization of the property of interest with respect to the latent space, and decode the result to produce an optimized molecule. We constrain optimization to the reparametrized $n - 1$ dimensional sphere of radius $\sqrt{n-1}$ for each n-dimensional layer of the hierarchical prior by optimizing the angle directly, since almost all of the probability mass of a Gaussian distribution lies in this thin spherical shell [25, Gaussian Annulus Theorem].

3 Results

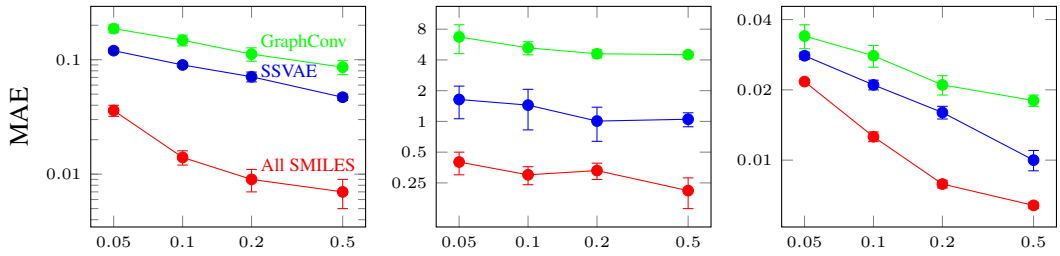
We evaluate the All SMILES VAE on standard 250,000 and 310,000 element subsets [6, 26] of the ZINC database of small organic molecules [27, 28]. We also evaluate on the Tox21 dataset [23, 24] in the DeepChem package [29], with binarized binding affinities of 7831 compounds against 12 proteins.

Using the approximating posterior as the encoder, but always selecting the mean of each conditional Gaussian distribution, and using beam search over the conditional likelihood as the decoder, 87.4% $\pm$ 1% of a held-out test set of ZINC250k (80/10/10 train/val/test split) is reconstructed accurately. With the same beam search decoder, 98.5% $\pm$ 0.1% of samples from the prior decode to valid SMILES strings. All molecules decoded from a set of 50,000 independent samples from the prior were unique, 99.958% were novel relative to the training dataset, and their average synthetic accessibility score [30] was 2.97 ± 0.01, compared to 3.05 in the ZINC250k dataset used for training.

As Figure 4 demonstrates, we significantly improve the state-of-the-art in the semi-supervised prediction of simple molecular properties, including the log octanol-water partition coefficient (logP), molecular weight (MW), and quantitative estimate of drug-likeness (QED) [22]. We achieve a similar improvement in fully supervised property prediction, as shown in Table 2. We also surpass the state-of-the-art in toxicity prediction on the Tox21 dataset [23, 24], obtaining an AUC-ROC of 0.875 $\pm$ 0.0008, as shown in Table 2. We refrain from ensembling our model, or engineering features using expert chemistry knowledge, as in the previous state-of-the-art method achieving an AUC-ROC of 0.862 [31].

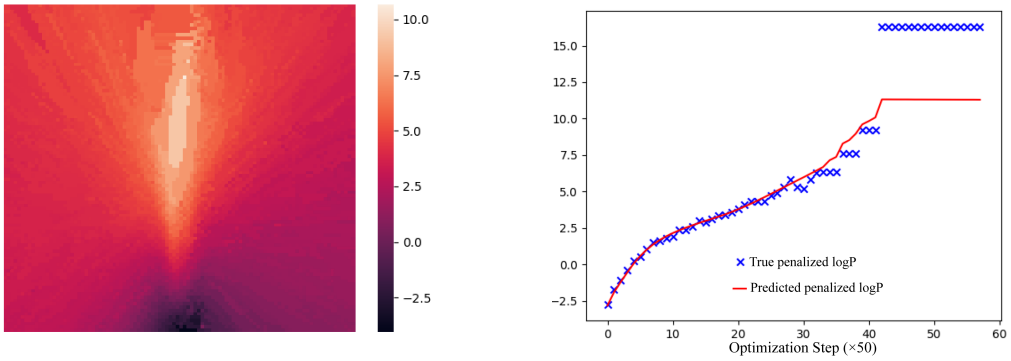
Accurate property prediction only facilitates effective optimization if the true property value is smooth with respect to the latent space. In Figure 5a, we plot the true (not predicted) logP over a densely sampled 2D slice of the latent space, where the y axis is aligned with the logP linear regressor.

We maximize the output of our linear and logistic property regressors, plus a log-prior regularizer, with respect to the latent space, subject to a hierarchical radius constraint. After optimizing in the



(a) LogP (b) MW (c) QED

Figure 4: Semi-supervised mean absolute error (MAE) ± the standard deviation across ten replicates for the log octanol-water partition coefficient (a), molecular weight (b), and the quantitative estimate drug-likeness [22] (c) on the ZINC310k dataset. Plots are log-log; the All SMILES MAE is a fraction of that of the SSVAE [26] and graph convolutions [32]. Semi-supervised VAE (SSVAE) and graph convolution results are those reported by Kang et al. [26].



(a) True logP over a 2D slice of latent space (b) Predicted and true logP over optimization

Figure 5: Dense decodings of true logP along a local 2D sheet in latent space, with the y axis aligned with the regressor (a), and predicted and true penalized logP across steps of optimization (b).

latent space with ADAM [33], we project back to a SMILES representation of a molecule with the decoder. Following prior work, we optimize QED and normalized logP penalized by the synthetic accessibility score and the number of large rings [7–9, 11, 34, 35]. Figure 5b depicts the predicted and true logP over an optimization trajectory, while Table 1 compares the top three values found amongst 100 such trajectories to the previous state-of-the-art.[1] The molecules realizing these property values are shown in Figure 7 of the Supplementary materials.

Table 1: Properties of the top three optimized molecules trained on ZINC250k.

MODEL	PENALIZED LOGP
JT-VAE [11]	5.30, 4.93, 4.49
GCPN [34]	7.98, 7.85, 7.80
MOLDQN [35]	8.93, 8.93, 8.91
ALL SMILES	12.31, 12.13, 12.01
All SMILES (KL unscaled)	**29.80, 29.76, 29.11**

MODEL	QED
JT-VAE [11]	0.925, 0.911, 0.910
CGVAE [12]	0.938, 0.931, 0.880
GCPN [34]	0.948, 0.947, 0.946
MolDQN [35]	**0.948, 0.948, 0.948**
All SMILES	**0.948, 0.948, 0.948**

[1]Zhou et al. [35] appear to report unnormalized penalized logP values: 11.84, 11.84, 11.82. In Table 1, we recompute normalized values for their best molecules. Recently, Winter et al. [17] reported molecules with penalized logP as large as 26.1, but train on an enormous, non-standard dataset of 72 million compounds aggregated from the ZINC15 and PubChem databases.

References

[1] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, and A. Aspuru-Guzik, "What is high-throughput virtual screening? A perspective from organic materials discovery," *Annual Review of Materials Research*, vol. 45, pp. 195–216, 2015.

[2] D. Stumpfe and J. Bajorath, "Exploring activity cliffs in medicinal chemistry: miniperspective," *Journal of medicinal chemistry*, vol. 55, no. 7, pp. 2932–2942, 2012.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, pp. 1278–1286, 2014.

[5] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[6] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[7] H. Dai, Y. Tian, B. Dai, S. Skiena, and L. Song, "Syntax-directed variational autoencoder for structured data," *arXiv preprint arXiv:1802.08786*, 2018.

[8] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," *arXiv preprint arXiv:1703.01925*, 2017.

[9] B. Samanta, A. De, G. Jana, P. K. Chattaraj, N. Ganguly, and M. Gomez-Rodriguez, "NeVAE: a deep generative model for molecular graphs," *arXiv preprint arXiv:1802.05283*, 2018.

[10] T. Aumentado-Armstrong, "Latent molecular optimization for targeted therapeutic design," *arXiv preprint arXiv:1809.02032*, 2018.

[11] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," *arXiv preprint arXiv:1802.04364*, 2018.

[12] Q. Liu, M. Allamanis, M. Brockschmidt, and A. L. Gaunt, "Constrained graph variational autoencoders for molecule design," *arXiv preprint arXiv:1805.09076*, 2018.

[13] J. Mueller, D. Gifford, and T. Jaakkola, "Sequence to better sequence: continuous revision of combinatorial structures," in *International Conference on Machine Learning*, pp. 2536–2544, 2017.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[15] E. Bjerrum and B. Sattarov, "Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders," *Biomolecules*, vol. 8, no. 4, p. 131, 2018.

[16] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert, "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations," *Chemical Science*, 2019.

[17] R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noe, and D. Clevert, "Efficient multi-objective molecular optimization in a continuous latent space," 2019.

[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[19] J. T. Rolfe, "Discrete variational autoencoders," *arXiv preprint arXiv:1609.02200*, 2016.

[20] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Advances in neural information processing systems*, pp. 1945–1953, 2017.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature chemistry*, vol. 4, no. 2, p. 90, 2012.

[23] R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. A. Shahane, A. Rossoshek, and A. Simeonov, "Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs," *Frontiers in Environmental Science*, vol. 3, p. 85, 2016.

[24] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "Deeptox: toxicity prediction using deep learning," *Frontiers in Environmental Science*, vol. 3, p. 80, 2016.

[25] A. Blum, J. Hopcroft, and R. Kannan, *Foundations of Data Science*. June 2017.

[26] S. Kang and K. Cho, "Conditional molecular design with deep generative models," *arXiv preprint arXiv:1805.00108*, 2018.

[27] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "Zinc: a free tool to discover chemistry for biology," *Journal of chemical information and modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.

[28] T. Sterling and J. J. Irwin, "Zinc 15–ligand discovery for everyone," *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.

[29] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[30] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," *Journal of cheminformatics*, vol. 1, no. 1, p. 8, 2009.

[31] M. Zaslavskiy, S. Jégou, E. W. Tramel, and G. Wainrib, "Toxicblend: Virtual screening of toxic compounds with ensemble predictors," *Computational Toxicology*, vol. 10, pp. 81–88, 2019.

[32] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of computer-aided molecular design*, vol. 30, no. 8, pp. 595–608, 2016.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] J. You, B. Liu, R. Ying, V. Pande, and J. Leskovec, "Graph convolutional policy network for goal-directed molecular graph generation," *arXiv preprint arXiv:1806.02473*, 2018.

[35] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of molecules via deep reinforcement learning," *arXiv preprint arXiv:1810.08678*, 2018.

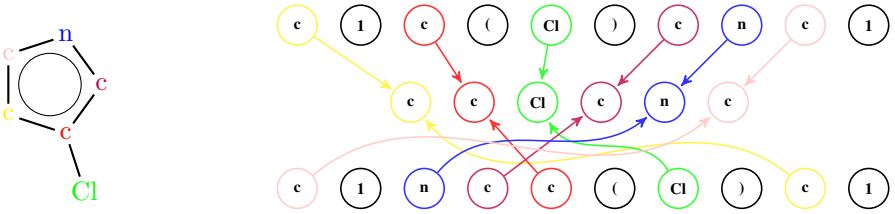
[36] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[37] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, pp. 2224–2232, 2015.

[38] J. Li, D. Cai, and X. He, "Learning graph-level representation for drug discovery," *arXiv preprint arXiv:1709.03741*, 2017.

[39] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, and V. S. Pande, "Potentialnet for molecular property prediction," *ACS central science*, vol. 4, no. 11, pp. 1520–1530, 2018.

A Supplementary figures



(a) Original molecule (b) Pooling of two SMILES strings representing the same molecule

Figure 6: To pass information between distinct paths implicit in multiple SMILES representations of a molecule, the encoder pools the representation of each atom across multiple SMILES strings.

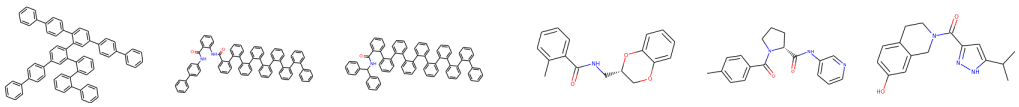
Table 2: Fully supervised regression on ZINC250k (a), evaluated using the mean absolute error; and Tox21 (b), evaluated with the area under the receiver operating characteristic curve (AUC-ROC), averaged over all 12 toxicity types. Aside from All SMILES, results in (a) are those reported in [6].

(a) ZINC250k

MODEL	MAE LOGP	MAE QED
ECFP [36]	0.38	0.045
CVAE [6]	0.15	0.054
CVAE ENC [6]	0.13	0.037
GRAPHCONV [37]	0.05	0.017
All SMILES	**0.005** $\pm$ **0.0006**	**0.0052** $\pm$ **0.0001**

(b) Tox21

MODEL	AUC-ROC
GRAPHCONV [29]	0.829 $\pm$ 0.006
LI, CAI, & HE [38]	0.854
POTENTIALNET [39]	0.857 $\pm$ 0.006
TOXICBLEND [31]	0.862
All SMILES	**0.875** $\pm$ **0.0008**



(a) Molecules with the top three penalized logP values (b) Molecules with the top three QED values

Figure 7: Molecules produced by gradient-based optimization in the All SMILES VAE.